SAND2020-4461C

Semiconductor Research Corporation®

Decadal Plan for Semiconductors:
New Compute Trajectories for Energy-Efficient Computing

Key Messages

Sandia National Laboratories
Livermore, CA
October 15-16, 2019

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

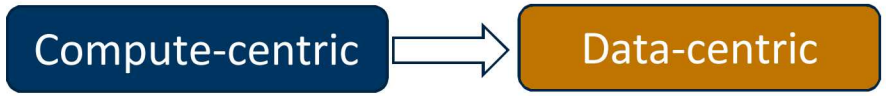
Session 1: ICT Systems: Fundamentals and Application Drivers

James Ang, PNNL
Roy Campbell, DoD
Jackie Chen, SNL
Stephen Kosonocky, AMD
John Owens, UC-Davis
Rob Aitken, ARM
David Wentzlaff, Princeton U

1

Zettascale computing may be the 1st milestone we will not be able to reach with the current technology

2



Transition from compute-centric to data-centric information processing

3 ML is a huge compute consumption

4 High-speed interfaces improve at much lower rate than computing

5 Monolithic 3D is coming

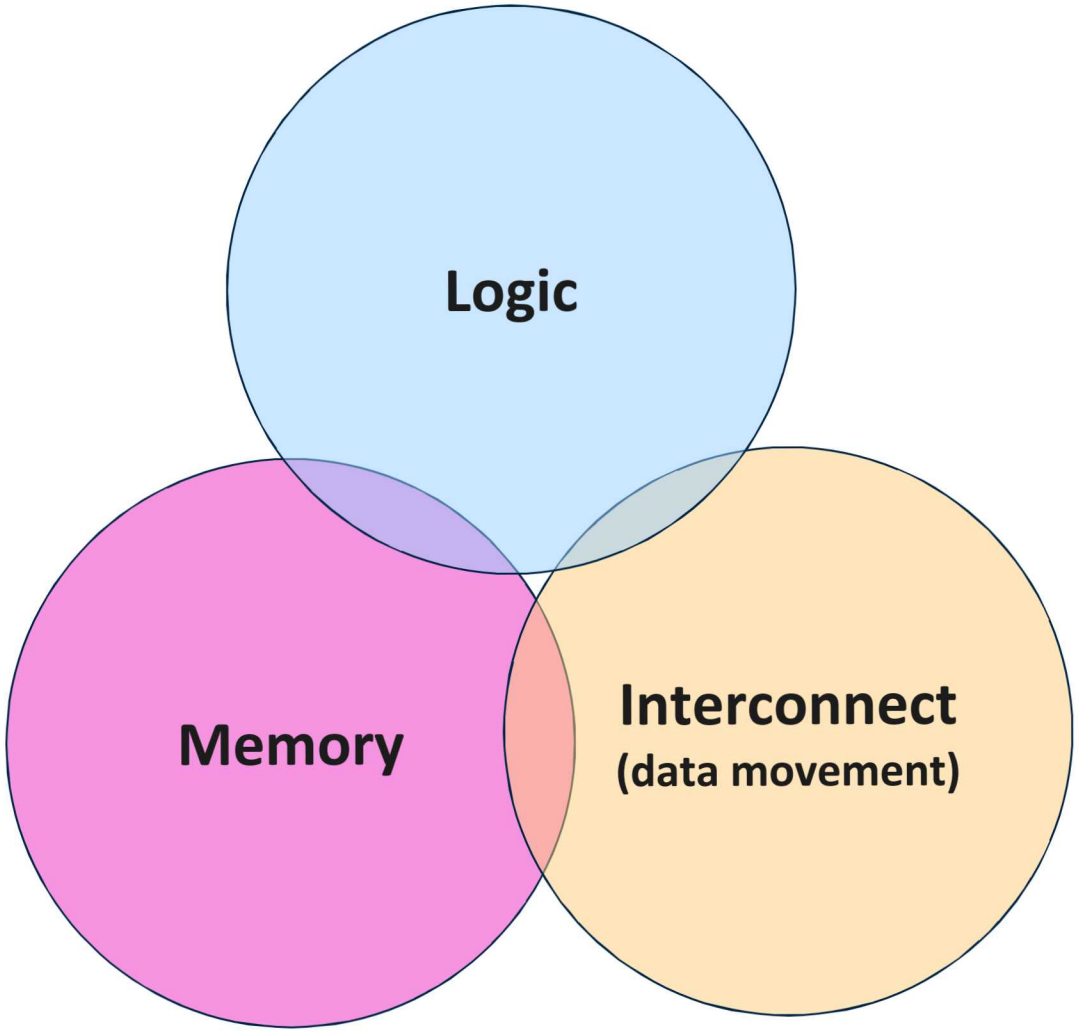
6 Computation is ‘free’, communication is expensive

7 Limits on accelerators!

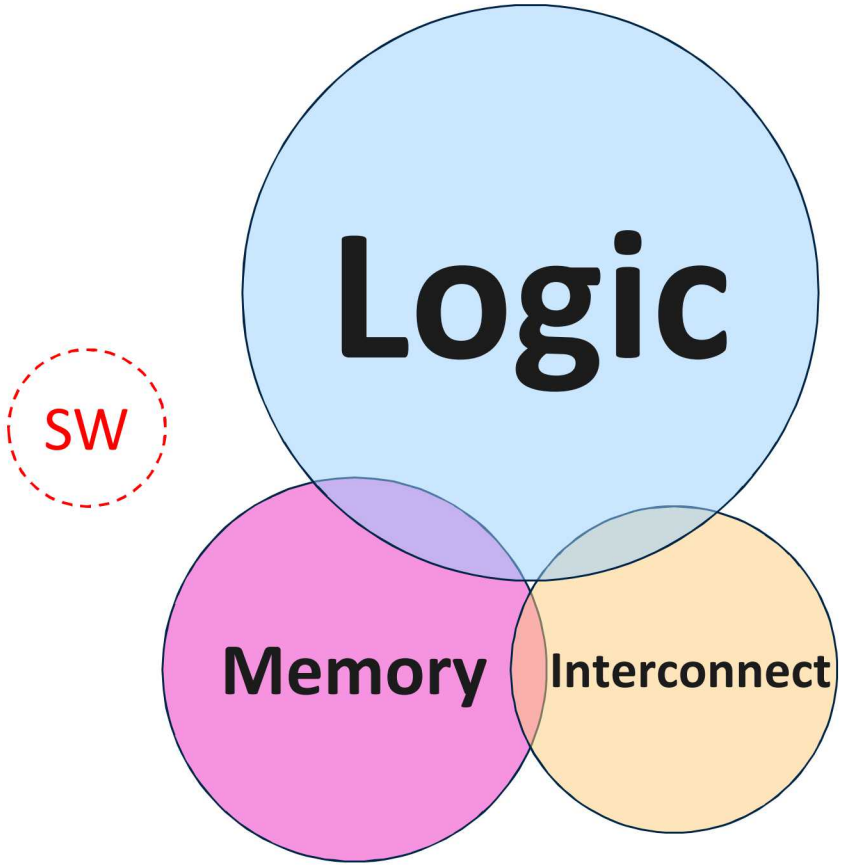
8 Software is King

9 Emphasis on SW-HW codesign

Three Cornerstones of Computing

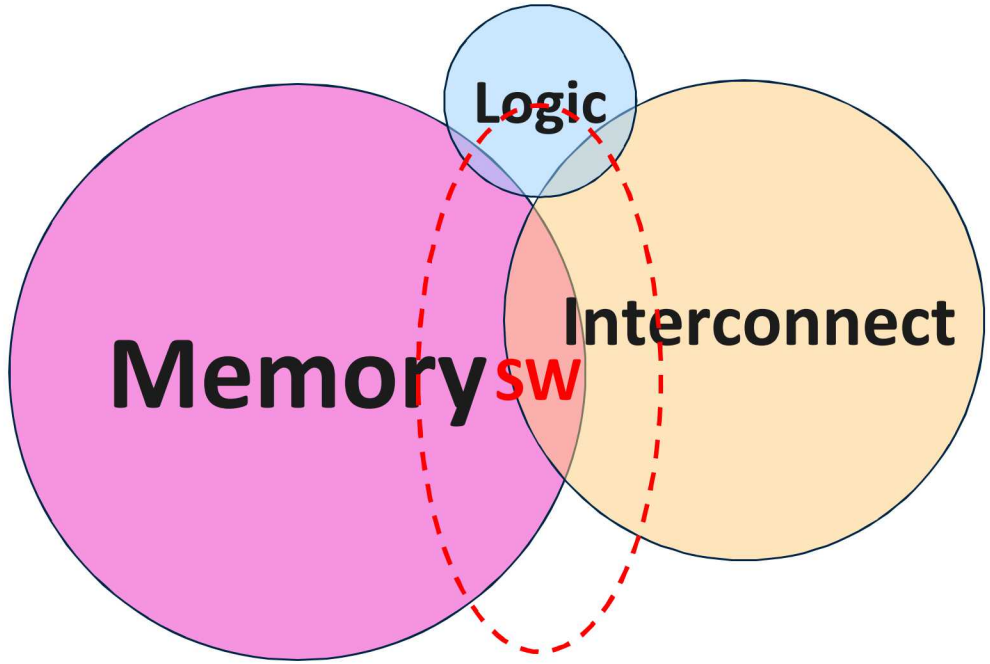


Three Cornerstones of Computing



1990

Computation is free!



Memory and communication are expensive…

Emphasis on SW-HW codesign

2019

Session 2: Impact of emerging device technologies

Ian Young, Intel
Naresh Shanghag, U Illinois
Kaushik Roy, Purdue U
Yichen Shen, Lightintelligence
Steve Trimberger, U Maryland
Dmitri Nikonov, Intel
Rob Clark, TEL

1 Turing → Shannon → Approximate computing

Shannon framework for computation

2 Spintronic devices are compatible with the Shannon framework

p-bits

3 NVM + CPU

Integration nonvolatile memory in compute node holds promise, but requires a lot of new efforts

4 Analog photonic computing

5 Cryogenic computing

6 Compute in interconnects

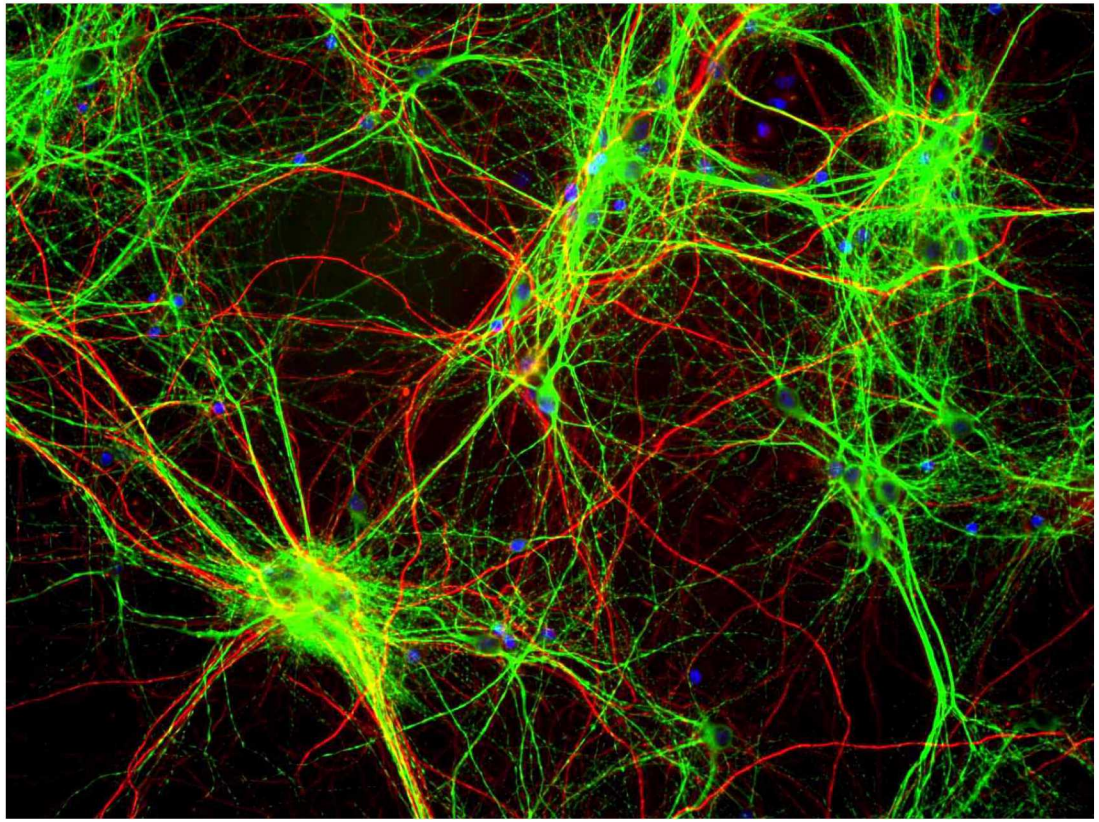
7 Compute in memory

8 Moore's law doesn't end with silicon!

Brain computes BOTH with interconnects and with memory

In the human brain, the distribution of **Ca** ions in dendrites represents a crucial variable for processing and storing information.

Ca ions enter the dendrites through voltage-gated channels in a membrane, and this leads to rapid local modulations of calcium concentration within dendritic tree



DENDRITES ARE LIKE MINI-COMPUTERS IN YOUR BRAIN

Source: FUTURITY

S. L. Smith et al, "Dendritic spikes enhance stimulus selectivity in cortical neurons in vivo", Nature 503 (2013) 115

C. Koch, "Computation and single neuron", Nature 385 (1997) 207

Session 3: Brain-inspired computing

Titash Rakshit, Samsung
Bruno Olshausen, UC Berkeley
Brad Aimone, Sandia
Stefano Ambrogio, IBM
Aaron Voelker, Applied Brain Res
Don Norman, UC San Diego

1. Spiking facilitates developing algorithms that more directly leverage time in computing
2. Nonlinear processing in dendritic trees (compute in interconnects)
3. Spiking + Analog
   We can do both!
4. Information should be naturally 'spike-based'
5. High-dimensional representation
6. Supercompression of information: $1:10^5$
7. Brain = Electrical + Chemical
   Need for 'wetware'?
8. Injection 'emotions' in computation?

Session 4: AI Engines

Heike Riel, IBM
Fred Streitz, DoE
Steven Lee, DoE
Anand Raghunathan, Purdue U
Tayfun Gokmen, IBM
Shimeng Yu, GeorgiaTech

1. AI accelerators/processors: "*Cambrian Explosion*" needs to happen!
2. Need: AI hardware roadmap
3. Ternary DNNs
4. Compute in memory
5. Needed: Edge AI
6. Analog vector-matrix multiplication
7. Rapid increase in model sizes leads to memory capacity and bandwidth demand
8. AI needs more compute!
9. Need: More 'intelligent' AI (beyond pattern recognition)
10. Can 'general purpose' AI hardware in the future be more energy efficient on a global scale than CPU-based systems?

Session 5: Large-scale Quantum Computing

Rafic Makki, Mubadala
Chris Monroe, U Maryland
Chad Rigetti, Rigetti Comp
Michael Biercuk, U Sydney
Oliver Dial, IBM
Edoardo Charbon, EPFL

1. QC decouples compute power from energy consumption ?
2. Quantum cloud service is a reality
3. 1st application of QC: Quantum Chemistry
4. Current status: 72 qubit system (Google)
5. Needed: error-corrected qubits
6. Electronic interfaces for quantum processors